An Empirical Study on the Perceived Fairness of Realistic, Imperfect Machine Learning Models

Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, Blase Ur
University of Chicago
{harrisong,jchanson,jacinto,julioramirez,blase}@uchicago.edu

ABSTRACT

There are many competing definitions of what statistical properties make a machine learning model fair. Unfortunately, research has shown that some key properties are mutually exclusive. Realistic models are thus necessarily imperfect, choosing one side of a trade-off or the other. To gauge perceptions of the fairness of such realistic, imperfect models, we conducted a between-subjects experiment with 502 Mechanical Turk workers. Each participant compared two models for deciding whether to grant bail to criminal defendants. The first model equalized one potentially desirable model property, with the other property varying across racial groups. The second model did the opposite. We tested pairwise trade-offs between the following four properties: accuracy; false positive rate; outcomes; and the consideration of race. We also varied which racial group the model disadvantaged. We observed a preference among participants for equalizing the false positive rate between groups over equalizing accuracy. Nonetheless, no preferences were overwhelming, and both sides of each trade-off we tested were strongly preferred by a non-trivial fraction of participants. We observed nuanced distinctions between participants considering a model "unbiased" and considering it "fair." Furthermore, even when a model within a trade-off pair was seen as fair and unbiased by a majority of participants, we did not observe consensus that a machine learning model was preferable to a human judge. Our results highlight challenges for building machine learning models that are perceived as fair and broadly acceptable in realistic situations.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in interaction design*; • **Computing methodologies** → *Machine learning*;

KEYWORDS

Fairness, Accountability, Machine Learning, Survey, Data Science

ACM Reference Format:
Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, Blase Ur. 2020. An Empirical Study on the Perceived Fairness of Realistic, Imperfect Machine Learning Models. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3351095.3372831

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02.
https://doi.org/10.1145/3351095.3372831

1 INTRODUCTION

As machine learning (ML) is used to automate increasingly significant decisions, such as predicting criminal defendants' risk of recidivism [7], activists and journalists have raised the alarm about issues of bias and a lack of fairness [2]. Researchers have responded by attempting to define fairness mathematically. There are dozens of competing definitions that embed different notions of what is fair [19]. Recent work has noted that these fairness definitions address a relatively narrow set of considerations [25]. In light of this, some have suggested that these definitions can be deployed to understand and address specific problems [22]. Unfortunately, some key definitions of fairness are incompatible [5, 17]. Therefore, in reality, building models that embed some of these definitions will require making difficult trade-offs between competing notions of fairness. We conducted an empirical user study to understand attitudes about these sorts of potentially difficult trade-offs.

Consider the well-worn example of COMPAS. In 2016, ProPublica reported that COMPAS, an ML tool used to assess criminal defendants' fitness for parole or bail, had a false positive rate for black defendants nearly twice that for white defendants [2]. The tool's makers responded by noting that the accuracy of the tool was equalized between the two racial groups and that the tool was therefore not biased [7]. Subsequent work has shown that one cannot achieve both conditions simultaneously [5, 17]. In other words, there was a necessary trade-off between equalizing either false positive rates or accuracy across racial groups. It is not clear whether the makers of COMPAS were aware of the disparate false positive rates prior to ProPublica's reporting, so it is not clear if preferring equalized accuracy was a conscious decision. Regardless, COMPAS implicitly embeds the notion that equalizing accuracy, rather than the false positive rate, is the correct definition of fairness.

A number of studies in recent years have aimed to quantify attitudes about the fairness of particular ML models [4, 9, 14, 15, 26]. However, few studies investigate attitudes about the difficult choices and fairness-related trade-offs inherent in realistic applications of ML. Understanding attitudes about these trade-offs can inform technical and regulatory interventions. Documenting what people outside of specialized fields think about these problems can help inform and facilitate future multi-stakeholder processes.

We conducted a survey-based experiment of 502 Mechanical Turk workers. In the context of making bail decisions, this experiment tested a fairness-related trade-off randomly assigned to each participant. Each pair captured one side of the type of trade-off that an ML developer might face when trying to make a model as fair as possible in realistic circumstances. For example, one set of participants was prompted to choose between a model that equalized accuracy at the expense of disparate false positive rates across racial groups, and one that did the opposite. We tested trade-offs

of all pairwise combination of four desirable fairness properties: equalized accuracy; equalized false positive rates; equalized outcomes; and explicitly excluding race as a model input. Each model in a pair had a disparity in one model property and equality in the other. Through this study, we investigated the following questions:

- **RQ1** When choosing between models exhibiting the two sides of a difficult trade-off, which do people prioritize?

We observed a statistically significant trend, but not full consensus, in participants prioritizing the equalization of false positive rates across groups rather than the equalization of accuracy. Notably, this preference is the opposite of what the COMPAS tool did, which may partially explain the controversy around the tool. For many other trade-offs, we saw a wide distribution in participants' preferences. We observed a non-trivial fraction of strong opinions preferring each side of each trade-off we tested, highlighting the difficulty of building ML models with broadly acceptable fairness characteristics in realistic circumstances.

- **RQ2** What models that encapsulate difficult, yet realistic, trade-offs do people perceive as fair or biased?

Participants tended toward rating most models we tested as not biased. However, participants tended to consider it biased when outcomes were equalized at the expense of disparate false positive rates that disadvantaged African-American defendants. Only one model we tested — one in which more African-American defendants than White defendants were granted bail even though race was not used as an input to the model — was overwhelmingly considered fair. In many other cases, opinions were mixed and fairly polarized.

A model's perceived lack of bias did not necessarily imply a perception of fairness. We observed a number of cases in which participants considered a model not to be biased, yet also did not consider it to be fair. In some cases, this was because of high false positive rates that were nonetheless equal across racial groups.

- **RQ3** Do people prefer to use an imperfect model or rely on a human judge?

For most trade-offs we investigated, we found a preference for a human judge over either ML model the participant saw. Even when the participant considered one or both of these models unbiased, they often still preferred a human judge. They justified this preference based on a human judge's accountability, capacity for ethical reasoning, and ability to make individualized decisions.

- **RQ4** To what extent do responses vary based on which racial group the model disadvantages?

For each trade-off, we randomly assigned whether White or African-American defendants would be disadvantaged by the disparate property. We observed some potential, yet not statistically significant, differences in the distribution of responses.

In sum, our empirical user study is a first step in unpacking how people view the bias and fairness of ML models encoding difficult trade-offs related to fairness.

2 RELATED WORK

We first summarize some of the most straightforward definitions of fairness, including those we investigate in this study (Sec. 2.1). Afterwards, we present prior studies that, like ours, empirically investigate humans' perceptions of the fairness of ML models (Sec. 2.2).

2.1 Fairness Definitions and Trade-offs

Researchers have proposed dozens of definitions of fairness for ML models [19]. Many of these definitions center on one or more properties of the model. In this section, we provide a brief overview of two classes of fairness definitions most relevant to our study. Definitions in the first class, *group fairness*, concern a model's comparative treatment of different groups (e.g., demographic groups). Definitions in the second class, *procedural fairness*, concern the process by which individuals are judged. While other classes of fairness definitions have been proposed, including definitions concerned with the treatment of individuals, all four definitions we test in our study fall into one of these two classes. We conclude this section by highlighting how some of these definitions are mutually exclusive, motivating our study of the trade-offs between definitions.

2.1.1 Group Fairness. Group fairness refers to the class of definitions that examine differences across groups, such as the different demographic groups represented in a dataset. These definitions frequently propose that some property, such as the model's accuracy [17] or false positive rate [5, 17], should be equal across groups. As highlighted in the COMPAS controversy, accuracy and false positive rates are two of the most prominent examples of model properties used to evaluate group fairness [5]. Another example of a group fairness definition is disparate impact, or demographic parity. Originating from employment discrimination law, disparate impact is the idea that a sufficiently large difference in favorable classification rates is rebuttable evidence of discrimination. In the ML context, disparate impact is frequently parameterized as attempting to equalize outcomes across groups [8].

More complex group fairness definitions have also been proposed. Under equalized odds, the true positive and false positive rates must be equalized between groups [12]. Equalized opportunity is less strict, requiring only the true positive rates be equalized [12]. Both equalized odds and opportunity encode the idea that fairness consists of being right at the same rate across groups. Refinements to these definitions often look at discrimination through a causal lens [16]. The intuition behind causal definitions is that the same distribution could be caused by a fair or an unfair social process.

In our study, we test conditions in which accuracy, false positives, and outcomes vary across groups (see Sec. 3). We do not test more complex definitions because they are difficult to explain succinctly.

2.1.2 Procedural Fairness. In contrast, procedural fairness focuses on how a model makes decisions. For instance, a procedural fairness approach might argue that a model that makes decisions using race as an input variable is always unfair, regardless of the model's outcome. The input variables most widely discussed in procedural fairness (e.g., race, gender, and age) are protected classes under employment discrimination law. Researchers have also proposed other ways of determining which variables are fair to use in ML models. Some have argued that for a model to be fair, it must give individuals the ability to change a model's decision [27]. Actionable variables might include employment status, while immutable variables would include race or the age of first arrest. Others have proposed the notion of process fairness in which fair variables are identified by surveying the public [11].

In our experiment, we vary whether or not race is included as an explicit model input. The explicit consideration of a data subject's race would violate almost all procedural fairness definitions. While process fairness considers the explicit use of variables, unfair variables may be implicitly encoded in data through combinations of seemingly fair variables. A large literature proposes techniques for auditing models for such correlations [1, 6, 21, 23].

2.1.3 Tensions Between Fairness Definitions. Unfortunately, certain key fairness definitions are mutually exclusive or otherwise incompatible. Kleinberg et al. [17] and Chouldechova et al. [5] both independently argue that it is only possible to achieve equally accurate risk scores across groups and equally balanced risk quantiles across groups under very specific conditions. Grgić-Hlača et al. find that one can sometimes achieve both process fairness and outcome fairness, but at an accuracy cost [11]. Furthermore, the use of race may be necessary during model development to audit the model or to achieve group fairness [29]. Indeed, the lack of data labeled with unfair attributes can prevent the analysis of fairness in practice [13].

Because many definitions of fairness can be mutually exclusive, realistic ML models often cannot simultaneously satisfy all desired definitions. In light of these inherent trade-offs, we design the models in our survey-based experiment to be imperfect, satisfying one definition of fairness, yet violating another.

2.2 Empirical Studies of Fairness

Researchers have begun to investigate humans' attitudes toward ML systems. Qualitative studies have found that people from marginalized communities do have experiences with algorithmic discrimination, though they do not always use that term for it [28]. Similar to us, a handful of empirical studies have collected attitudes about specific model properties. For example, Srivastava et al. ask participants to choose between a succession of pairs of models to identify which group fairness definition best captures people's perceptions of fairness [26]. Through twenty comparisons generated through an adaptive algorithm, they converge upon a given respondent's preferred notion of fairness. In both risk prediction and medical diagnostic contexts, their participants prefer demographic parity (equality across groups in the percentage predicted to receive a positive classification) over other definitions. As we discuss in Section 5.2, using a different protocol we come to a different conclusion.

Within process fairness, Grgić-Hlača et al. investigate attitudes about what features people consider fair to use [10]. In follow-up work, Grgić-Hlača et al. investigate why people consider those features fair or not [9]. They find a feature's perceived relevance, reliability, and volitionality drive assessments. They also find that support for the consideration of race increases from 17% to 42% when participants are told the use of race increases accuracy.

Two additional empirical studies focus on questions complementary to ours. In a loan-allocation context, Saxena et al. quantify attitudes about individual fairness [24]. They ask participants to rate the fairness of three different ways a loan officer could divide $50,000 between two individuals with different repayment rates (and in one iteration, different races). Participants rated giving the entire sum to the candidate with the higher repayment rate as more fair than dividing it equally only when the candidate with the higher rate was black. Kennedy et al. investigate the relationship between trust and model properties. They use an experiment in which participants choose between pairs of risk assessment models [15]. The models vary at random in overall rates (e.g., true/false positives/negatives), the size of the training data, the number of features, the weight of features, the algorithm source, and more. While they investigate trust, they do not investigate fairness or test whether differences in model properties across groups affect trust. Awad et al. use a viral game to study variations across cultures in how participants would solve a variation of the trolley problem involving different demographic characteristics of the parties [3]. Our work is distinct from Awad et al. both in that we seek to understand a different form of decision, and that we do not attempt to make any kind of generalized claims about universal moral characteristics.

3 METHODOLOGY

To investigate perceptions of fairness in imperfect, yet realistic, ML models, we conducted an online, between-subjects, survey-based experiment. We asked participants to rate the fairness, bias, and utility of two models that exhibited both sides of a specified trade-off between two fairness-related model properties randomly selected from among the following four: accuracy, false positive rate, outcomes, and the explicit consideration of race. We graphically presented the properties encapsulated in this trade-off. Participants then chose which of the two models they preferred overall, as well as whether they preferred a human judge to either ML model.

On Amazon's Mechanical Turk, we recruited workers 18+ years old and located within the United States. We limited recruitment to workers with a 95%+ rating over 500+ HITs. We paid $2.50 for the survey, which took a median time of 14 minutes. We excluded data from participants whose free responses were off-topic or nonsensical (e.g., discussing the Tesla Model X car in their response). Exclusion happened after data collection, and all participants were paid regardless of whether we excluded their data from analysis. Our Institutional Review Board approved this experiment.

We used Mechanical Turk for participant recruitment because it provided a means of reaching participants who were largely non-technical and were unlikely to have formally engaged with academic debates about fair machine learning. We do not claim that this participant pool is representative of some larger public. Our participant pool largely did not have technical backgrounds.

Below, we detail the structure of the survey (Sec. 3.1) and the specific trade-offs we tested (Sec. 3.1.1). We then describe our quantitative (Sec. 3.2) and qualitative (Sec. 3.3) analyses. The supplementary appendix includes the full survey instrument.

3.1 Survey Structure

To familiarize participants with the topic of machine learning, the survey began by giving a high-level description of how ML models can be used to make predictions. We then told them a city was considering using an ML model to decide whether to grant bail to defendants charged with non-violent crimes.

Each participant was randomly assigned a pair of models exemplifying a trade-off between two fairness-related model properties (see Sec. 3.1.1). Each model satisfied one definition of fairness, but violated the other. We presented these properties visually (see Fig. 1).

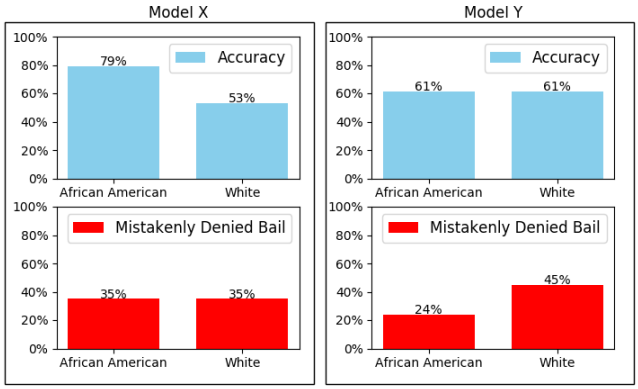

Figure 1: The comparison presented to participants in the condition testing accuracy and false positives (FP-Acc).

Property	Disparate Rates	Equalized Rate
Accuracy (**Acc**)	79% vs. 53%	61%
Outcomes	65% vs. 41%	53%
False Positives (**FP**)	45% vs. 24%	35%
Race usage	Is a model input	Is not a model input

Table 1: The properties investigated. The equal accuracy rate, as well as disparate rates for outcomes and false positives, were based on the COMPAS dataset [18].

We first presented the models individually, in randomized order. Participants rated the fairness, bias, and utility of each on a five-point Likert scale. Afterwards, we presented the two models side-by-side (as in Figure 1), asking participants to rate on five-point Likert scales which was more fair, biased, or useful. Also on five-point Likert scales, we asked participants to select which model they would prefer to see implemented, as well as whether they would prefer a human judge to their choice of either model.

3.1.1 Model Properties and Trade-offs. Each participant evaluated and compared two models that were imperfect in opposite ways. We randomly assigned each participant a pair of models representing the trade-off between two of the following four properties: *accuracy*, *false positives*, *outcomes*, and explicit *race usage*. We chose these properties because they have been widely discussed and can be expressed succinctly. In contrast, we chose not to test more complex definitions like equalized odds, which requires equalizing both the true and false positive rates [12]. Table 1 summarizes these properties, and their associated disparate and equalized rates. We tried to capture realistic trade-offs by using rates taken from ProPublica's analysis of COMPAS data [18]. Because some of our properties are counter-factuals, we fabricated the unknown rates (disparate accuracy and equalized false positive and outcome rates).

Below, we list the properties (and, in italics, the abbreviations we use throughout the paper) as we explained them to participants:

- **Accuracy (*Acc*):** "The accuracy of the model is the rate at which the model makes correct predictions. A prediction is correct if the model either predicts that a defendant will show up for their trial and they would, or that they will not show up for their trial and they would not have."
- **Outcomes:** "The probability of bail is the likelihood of a defendant being granted bail if the model is used."
- **False positive rate (*FP*):** "A defendant who is mistakenly denied bail is one that the model predicts would not show up for their trial when they would have." We explained false positives without using the term "false positives" to avoid confusion about what a "positive" classification meant.
- **Race usage:** If the model does not use race: "The model makes decisions with no knowledge of the race of the subjects. Other features like type of offense and number of previous offenses are used as input to the model." If the model uses race: "The model makes decisions with knowledge of the race of the subjects. Other features like type of offense and number of previous offenses are used as input to the model." We considered a model "equalized" if it did not use race as an explicit model input. We considered a model to be "disparate" if it did use race. This corresponds to the notion that considering race is generally undesirable.

For each participant, we randomly selected one of the six possible pairwise combinations of these four properties. In all models, we showed how the two properties varied across two subgroups: White defendants and African-American defendants. We also randomly assigned whether African-Americans were disadvantaged or whether White defendants were disadvantaged in the disparate rates, doubling the total number of conditions to twelve.

3.1.2 Terminology. As we present our results, we refer to the trade-off participants saw using the abbreviated names of the two properties involved, as well as the *disadvantaged* group. For example, "FP-Outcome-Maj" refers to the trade-off between false positives and outcomes in which the majority group (White defendants) is disadvantaged. At some points, we need to refer to the particular model within the pair. We use = and ≠ to indicate the equalized and disparate property, respectively. For example, the FP-Outcome-Maj condition includes the model = Outcome, ≠ FP, Maj and the model ≠ Outcome, = FP, Maj. When we quote participants, we identify them by participant number and condition (trade-off).

3.2 Quantitative Analysis

We mapped answers on five-point Likert scales to (-2, -1, 0, 1, 2). For example, participants could rate whether they would definitely (-2) or probably (-1) prefer a model, that they were unsure (0), or that they would probably (1) or definitely (2) prefer a human judge. For such answers, we tested whether participants' answers tended toward one answer (model), the other (human judge), or neither. We did so using the unpaired Wilcoxon Signed-Rank test, which tests whether a distribution is skewed around zero. Significance indicates answers tended toward one answer or the other.

For questions where participants individually rated each model (e.g., on fairness), we tested whether they tended to rate one model higher than the other. As each participant rated both models, the data was not independent. We thus used the paired Wilcoxon Signed-Rank test, which measures whether the distribution of differences between pairs of ratings is symmetric. Significance indicates one model was seen as more fair, biased, or useful than the other.

For each trade-off pair (e.g., FP-Outcomes), some participants saw a version where White defendants were disadvantaged (-Maj), while others saw a version where African-American defendants were disadvantaged (-Min). We compared the -Maj and the -Min versions of each pair using the Mann-Whitney U test (a non-parametric analog of the ANOVA test for comparing two groups).

For each of the above families of tests, we corrected for multiple testing with the Benjamini-Hochberg method.

As both fairness and bias ratings were ordinal, Likert-scale data, we calculated the correlation of these ratings with Kendall's τ.

3.3 Qualitative Analysis

We thematically coded free-response explanations participants gave for their choice of one model over another, their preference for a human judge over either model, and their ratings of fairness and bias. Two coders collaboratively developed a codebook from a sample of answers. Two coders then independently used that codebook to code the remaining answers. We allowed multiple codes per answer to capture the compositionality of responses.

We created three distinct codebooks. The first was for explanations of bias. It contained ten high-level codes, six of which had sub-codes. For this codebook, Cohen's $\kappa = 0.77$. The second was for why a participant chose one model over the other. It had seven codes (no sub-codes), and $\kappa = 0.71$. The last codebook assessed explanations of why humans or models were preferred. It had eight high-level codes, five of which had subcodes. For this codebook, $\kappa = 0.64$. The coders met to resolve disagreements.

We saw a wide variation across participants in the length of these responses. For example, free-response explanations of why the participant preferred one model over the other ranged in length from 2 to 79 words, with a median of 11. Most free-response questions had a similar distribution in terms of length.

4 RESULTS

We begin by describing our participants (Sec. 4.1). We detail which trade-offs participants preferred (Sec. 4.2) and whether they ultimately preferred a human judge (Sec. 4.3). We then unpack participants' ratings of fairness and bias, as well as how those concepts correlate (Sec. 4.4). Finally, we delve into the impact of varying which racial group was disadvantaged (Sec. 4.5).

4.1 Participant Demographics

We surveyed 502 individuals in a convenience sample from Mechanical Turk. Table 2 summarizes our participants' demographics, which we compare to those of a recent study by Redmiles et al. [20] comparing sampling methods. Our participant population skewed male. Consistent with Redmiles et al., participants were more highly educated than the overall population. Political affiliation was split; 48% of our participants described themselves as Democrats, 28% as Independent, and 20% as Republican. A large portion reported no experience with computer science (48%). A roughly similar percentage (41%) reported no experience with probability. Most respondents (77%) reported having no experience with machine learning.

In spite of this, most participants appeared to understand the mathematical concepts we presented to them. We tested participants' understanding of the graphs in our survey with a series of graph comprehension questions at the end of the survey. These graph understanding questions were imperfect and highly limited. However, a majority of participants were able to correctly identify valid and invalid inferences. Furthermore, we administered a cognitive reflection test (CRT), a series of three quantitative questions shown to correlate with quantitative reasoning ability. More than half of participants (54%) got all three questions correct, 17% got two, and 12% got one. The CRT was at the end of the survey and the answers are free-response integers. Participants' good CRT performance suggests both that they were taking the survey seriously and that they had decent quantitative reasoning skills.

		Ours	Redmiles et al.
Gender	Male	60%	50%
	Female	40%	48%
	Gender non-binary	<1%	—
Race/Ethnicity	White	76%	84%
	Black/African-American	10%	10%
	Hispanic/Latinx	5%	4%
	Other	—	5%
	Asian	9%	—
	Native American	2%	—
	Hawaiian/Pacific Islander	<1%	—
Income	$0-$49,999	52%	49%
	$50,000-$99,999	35%	38%
	$100,000+	11%	11%
Education	< High School	<1%	<1%
	High School	16%	12%
	Some college/Two-year degree	34%	41%
	Four-year degree or above	50%	46%

Table 2: Participants' demographics, which we compare to Redmiles et al.'s analysis of MTurk workers [20]. A dash indicates that the study did not use that particular category.

4.2 Preferred Trade-offs

We investigated six pairs of trade-offs between properties, each with variants in which majority (White) and minority (African-American) groups were disadvantaged. We observed a preference for equalizing false positives over equalizing accuracy. As shown in Figure 2, this preference was statistically significant, yet not overwhelming. Comparing false positives against accuracy, 53.7% of participants probably or definitely preferred the model that equalized false positive rates when White defendants were disadvantaged, and 56.8% of participants favored the model that equalized false positives when African-American defendants were disadvantaged. In contrast, only 14.6% and 18.2% of participants, respectively, probably or definitely preferred the model that equalized accuracy at the expense of disparate false positive rates. The rest were unsure.

That we observed this preference toward equalized false positive rates is particularly notable because ProPublica's reporting centered on COMPAS doing the opposite: equalizing accuracy at the expense of disparate false positive rates [2]. In short, this result highlights that COMPAS equalized the property that our participants preferred *significantly less*, potentially explaining some of the public controversy about the COMPAS system.

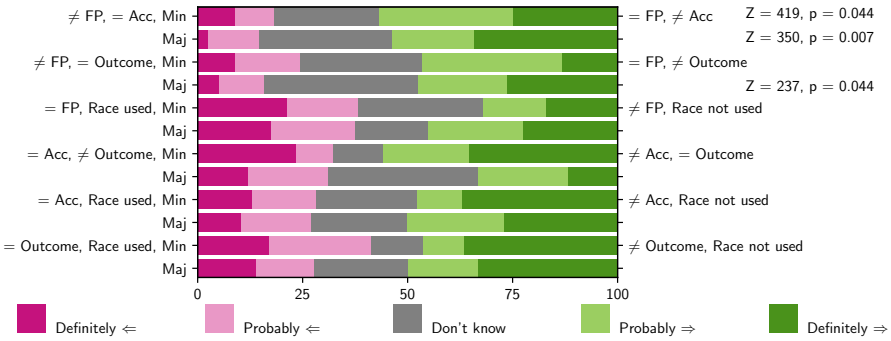

Figure 2: Participants' preferences for one model within a trade-off pair over the other. The left and right arrows indicate preferences toward the side of the trade-off listed on the left or right axes of the graph, respectively.

Among the remaining five trade-offs, again shown in Figure 2, we also observed a statistically significant preference toward equalizing false positives over equalizing outcomes when White defendants were disadvantaged (FP-Outcome-Maj). Here, 47.4% of participants preferred the model that equalized false positives, while only 15.8% preferred the model that equalized outcomes. This effect disappeared when African-Americans were disadvantaged.

Notably, for each trade-off, a non-trivial fraction of participants preferred *each side* of the trade-off (see Fig. 2). Recall that participants were assigned to one of twelve conditions (six trade-off pairs multiplied by two different groups that could be disadvantaged). For the nine conditions in which we did not observe a statistically significant preference, at least 24.4% of participants probably or definitely preferred each side of the trade-off. Furthermore, at least 8.9% of participants *definitely* preferred each side of the trade-off. In other words, a non-trivial fraction of participants would be unhappy no matter which side of the trade-off was chosen.

Our qualitative analysis of participants' free-text justifications for their choice emphasized that participants successfully identified the characteristics we were varying, yet did not shed much insight beyond this confirmation. When participants articulated specific reasons for their choice, they often did so by saying that equalizing the particularly property captured by the model they chose was more fair or less biased (34.1% of all explanations). For example, P452 (FP-Outcome-Min) said they chose the model that equalized outcomes because in the alternative model, "White people are unfairly being given bail more than blacks." An additional 26.3% of explanations were too vague to provide additional insight. For example, P42 (Acc-RaceUsage-Maj) wrote, "A combination of both can provide valuable information. Not including race changes the results significantly. Testing both models and determining how much each factor should weigh would help improve both models."

While few participants described fully what they meant by the terms, accuracy (mentioned in 18.9% of all answers) the use of race (13.3%), and equality (10.8%) were all invoked. Explanations invoking equality tended to further mention evenness and consistency. For example, P460 (FP-Outcome-Min) preferred to equalize outcomes because the "probability of bail is consistent across race."

When race usage was part of the trade-off, justifications often mentioned this property. For example, P245 (FP-RaceUsage-Min) preferred to equalize false positives even when explicitly considering race "because it is much more accurate overall." In contrast, P25 (Acc-RaceUsage-Maj) wrote, "I'd rather not have race as an issue for the computer to factor in because it is irrelevant to the decision being made." Participants who wrote that the model they saw was fair because it did not use race generally did not expand on what they thought "not using race" meant.

We intentionally limited our explanation for the use of race condition to a simple statement that race was not an explicit variable in the model, thus leaving the door open to participants identifying the issue of proxy variables. This reflects the perspective of process fairness, which is not explicitly concerned with whether some combination of variables can combine into another (impermissible) variable. Our qualitative responses indicate that participants largely seem to have assumed that "not using race" meant race did not factor into the classification at all. For the models where there was some disparity and race was not used, 6.2% of participants' qualitative responses pointed to the disparate quantity to indicate either that some other sensitive variable played a role in the model's predictions, or that the disparity indicated race was used. For example, one participant explained that a model that exhibited differences in outcomes, yet did not use race, was not at all fair by stating, "It states that race is not a factor, but clearly it is as it's showing that white have a higher probability of paying bail than African Americans" (P206, Outcome-RaceUse-Min).

4.3 Human Judges vs. Models

After participants saw both sides of the trade-off, we asked them whether they preferred a human judge or their choice of either of the two models they had seen. For eight out of the twelve conditions, we observed a statistically significant preference for human judges. Figure 3 shows this graphically. Among these eight conditions, the percentage of participants probably or definitely preferring a human ranged from 58.3% (Acc-RaceUsage-Maj) to 80.5% (FP-Acc-Maj).

We did not observe a statistically significant preference for using an ML model in any condition. The Outcome-RaceUsage conditions had the highest percent of participants who favored the model over a judge (44.4% for Outcome-RaceUsage-Maj and 41.5% for Outcome-RaceUsage-Min). Furthermore, these same conditions also had the highest proportion of people who strongly favored the model over a judge (19.4% and 19.5% for -Maj and -Min, respectively). That

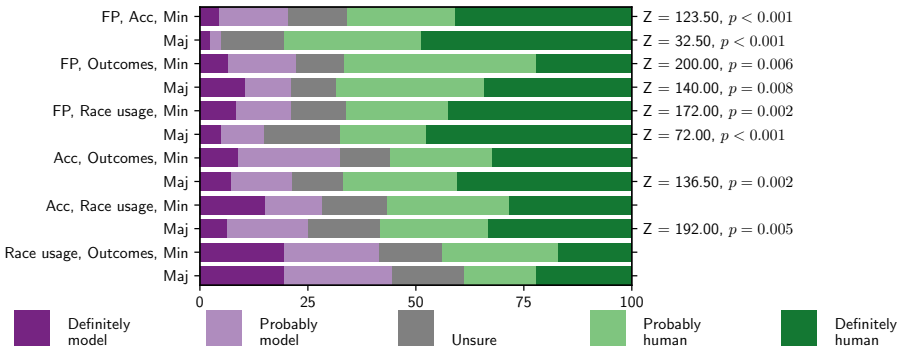

Figure 3: Participants' preferences for a human judge versus either model seen.

human judges were fallible and models were more objective appeared in 47% of participants' free-text justifications in these cases where the participant preferred a model over a judge. For example, P372 (Outcome-RaceUsage-Maj) wrote, "Human beings are biased and easily manipulated." Similarly, P392 (Outcome-RaceUsage-Min) wrote, "Models lack bias and personal experiences that they will not be able to use to make decisions." However, being able to use individualized judgment was also sometimes expressed positively.

In contrast, participants expressed a number of reasons for preferring a human judge. 76.7% of justifications mentioned a judge's ability to make individualized, case-by-case decisions. For example, P94 (FP-Acc-Maj) wrote, "I would use a human judge because I think that they would be able to see past the race of the offender and recognize outliers in their personalities and past that would make them more risky to not appear before trial." Similarly, P127 (FP-Acc-Maj) wrote, "A human judge would be able to take race out of the equation." An additional 29.1% of justifications discussed the judge as either more accountable or more ethical. Finally, 27.4% of justifications expressed that judges have ethical reasoning capabilities beyond those of models. According to P232 (FP-RaceUsage-Min), "I still think human judges can see things that a model cannot such as morals, values, and attitude. These cannot be taken into account by the model as it only looks at conditions known to it." Similarly, P384 (Outcome-RaceUsage-Min) wrote, "Programmers aren't trained to at least try not to be biased."

Even though participants preferred equalizing false positive rates to equalizing accuracy (as described in the previous section), they nonetheless still significantly preferred a human judge. In particular 65.9% of FP-Acc-Min participants and 80.5% of FP-Acc-Maj participants preferred a human judge to either model they saw.

4.4 Fairness and Bias

Prior to having participants compare the two models in their trade-off, we asked them to rate the fairness, bias, and usefulness of each model individually. Across most models, participants tended to lean toward rating the model as "not at all biased" or only "somewhat biased." For only one model did participants significantly lean toward rating it as biased. Nonetheless, participants tended not to rate these models as fair. In particular, for only one model did participants significantly lean toward rating it as fair. In contrast, for six others, they leaned towards rating it as "not at all fair" or only "somewhat fair." In spite of these trends, there were no models where participants' opinions were unanimous, and many where opinions were strongly divided.

4.4.1 Bias. Participants rated the bias of each of the two models they saw on a five-point scale from "completely biased" to "not at all biased." There were six trade-off pairs, two possible disadvantaged groups, and two models per pair, yielding twenty-four individual models. For sixteen of these twenty-four, participants significantly tended towards rating the model as not biased, as shown in Figure 4. For these sixteen models, the percentage of participants who rated the model as either "not at all biased" or only "somewhat biased" ranged from 56.3% to 73.9%. In contrast, for only one model (≠ FP, = Outcome, Min) were responses significantly skewed toward rating the model as biased. That is, participants tended to consider it biased when outcomes were equalized at the expense of disparate false positive rates that disadvantaged African-American defendants.

We also tested differences between bias ratings within a trade-off pair. Consistent with the results of the preferred side of trade-offs, we found a significant difference in bias assessments between models in the FP-Acc-Min trade-off ($W = 106$, $p = 0.034$). That is, participants felt that equalizing accuracy was more biased than equalizing false positive rates.

4.4.2 Fairness. As shown in Figure 4, we found fewer significant trends in participants' fairness ratings. For six of the twenty-four models, participants significantly tended toward rating the model as not fair. For these six models, between 63.6% and 87.8% of participants thought the model was "not at all fair" or only "somewhat fair." In contrast, for only one of the twenty-four models did we observe a significant trend toward considering a model fair. Overall, 80.6% of participants rated as "mostly fair" or "completely fair" the model in which more African-American defendants than White defendants were granted bail even though race was not used as an input to the model (≠ Outcome, Race not used, Maj). This perception of fairness did not persist when White defendants were favored (see Figure 4). Differences based on which group was disadvantaged are discussed further in Section 4.5. Like bias ratings, fairness ratings were polarized. 85% of participants rated at least one of the two models they saw as either "completely fair" or "not at all fair."

We also tested for differences in perceived fairness between models in a pair. The model equalizing false positives was rated as more fair than the model equalizing accuracy when Whites were

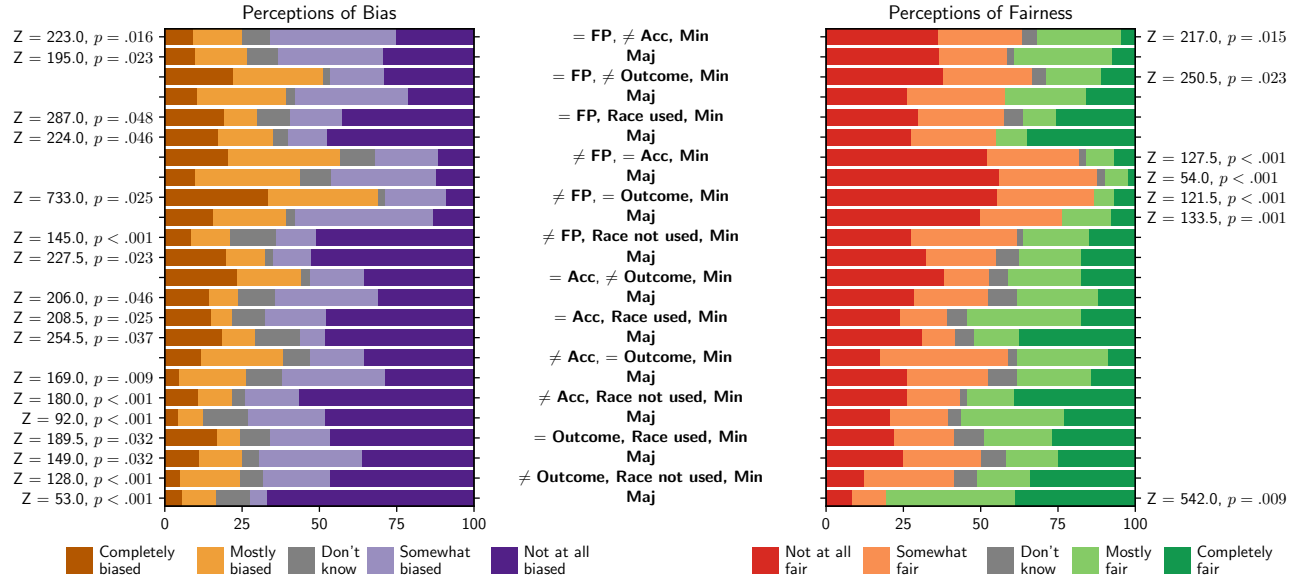

Figure 4: Participants' ratings of each model's bias (left) and fairness (right). The bold axis in the center indicates the condition.

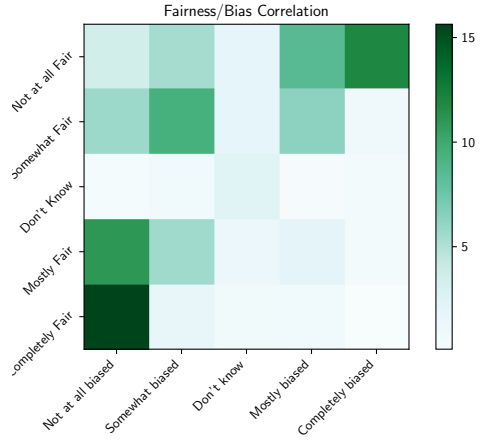

Figure 5: Correlation between ratings of fairness and ratings of bias for each model. Percentages are of total answers.

disadvantaged ($W = 289.5, p = 0.043$). This echoes a similar finding in differences of their bias ratings.

4.4.3 The Relationship Between Bias and Fairness. While one might assume that models that are not biased are thereby fair, we observed a much more complex and nuanced relationship between participants' ratings of bias and fairness. Participants tended to rate the model with equal outcomes and unequal false positive rates disadvantaging African-Americans (≠ FP, = Outcome, Min) as biased and not fair. Participants tended to rate the model where race was not used, but outcomes were unequal disadvantaging White defendants (≠ Outcomes, Race not used, Maj), as both fair and not biased. Participants tended to rate the model with equal false positives and unequal accuracy disadvantaging African-American defendants (= FP, ≠ Acc, Min) as not biased, yet also not fair.

We found that participants tended to rate models they considered biased as unfair. Graphically, the triangular nature of Figure 5 shows that models that were considered biased were almost never considered to be fair. The most common combination was that a model was "not at all biased" and also "completely fair." However, a model that they rated as not biased was not necessarily rated as fair. 39.0% of participants rated at least one model as either "not at all" or only "somewhat" biased, yet "not at all" or only "somewhat" fair. As shown in the leftmost column of Figure 5, some participants rated models as "not at all biased," yet only "mostly," "somewhat," or "not at all" fair. That is not to say that there is not a strong association between bias and fairness. The Kendall's τ correlation coefficient between fairness ratings and bias ratings was -0.513 ($p < .001$).

Ratings of a model being not biased, but also not fair, were most frequent in response to ≠ FP, = Outcome, Maj, composing 42.1% of all responses to that model. Similarly, they were 36.6% of all responses to the ≠ FP, = Acc, Maj model. Such answers tended to express ambivalence. For example, P447 (FP-Outcome-Maj) wrote that the model "seems somewhat fair but mistakenly denying bail to 35% of Whites and African Americans still seems like a high error rate." They then explained that they did not think the model was biased because "Model X mistakenly denies bail equally across both Whites and African Americans." In other words, high error rates could make a model unfair even though these rates were equal across racial groups (making the model unbiased).

4.4.4 The Relationship Between Bias and Judge Preferences. Initially, we had expected that participants who rated a model they saw as "not at all biased" would prefer a model over a human judge since they had an unbiased option. However, most of these participants nonetheless preferred a human judge over either model they saw. Figure 6 shows little difference in the distribution of preferences for a human judge versus a model between participants who rated both models they saw as unbiased and those who rated one model as biased (and the other as unbiased). About half of participants in

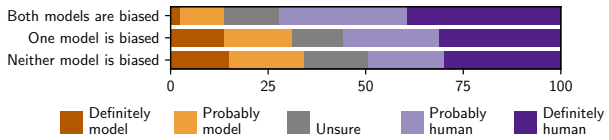

Figure 6: Preference for a human or a model, broken down by the participant's ratings of the individual models' bias.

each case "probably" or "definitely" preferred a human judge. Participants who reported both models they saw as at least somewhat biased were even more likely to prefer a human judge.

It is also noteworthy that even though 80.6% of participants thought that the ≠ Outcome, Race not used, Maj model was "mostly" or "completely" fair, only 44.4% of participants "probably" or "definitely" preferred a model instead of a human judge. This suggests that even relative consensus about the fairness of a model may be insufficient to produce consensus about whether an ML model is more appropriate to use than a human judge.

Qualitative coding of the reasons for bias ratings showed that participants largely understood the trade-offs, yet did not provide much deeper insight. Most frequently, participants reported finding a model biased because of disparate false positive rates (43.4% of responses), the explicit consideration of race (34.7%), disparate accuracy (33.5%), and disparate outcomes (29.7%).

4.4.5 Usefulness. Participants' ratings of a model's usefulness were largely redundant to those of bias and fairness. In the majority of models where we observed a statistically significant trend toward one side, the trends were toward the model not being useful. The one model where there was a statistically significant trend toward a model being useful was the already discussed ≠ Outcome-Race not used-Maj model. The results for models' perceived usefulness are shown in the supplementary appendix.

4.5 The Impact of Who Was Disadvantaged

For each of the six trade-offs, half of participants saw a model where the majority group (Whites) were disadvantaged by the disparate rate, whereas the other half saw a model where the minority group (African-Americans) were disadvantaged. We observed several instances in which there appear to be differences in participants' reactions between the -Maj and -Min variants. Despite these apparent differences, statistical testing did not distinguish between the -Maj and -Min variants. The model in which outcomes were equalized at the expense of disparate false positive rates was rated as "mostly biased" or "completely biased" by 68.9% of participants when African-Americans were disadvantaged, yet only by 39.5% of participants when Whites were disadvantaged. Similarly, in the model where race was not considered and outcomes were disparate across groups, 80.6% of participants rated the model as "mostly fair" or "completely fair" when Whites were disadvantaged, yet only 51.2% did so when African-Americans were disadvantaged. Though not statistically significant, the differences we observed suggest a more targeted study may have found significant differences. In particular, our statistical power in comparing the -Maj and -Min variants was relatively low given that each condition had roughly 20 participants. Furthermore, we were also making many comparisons and therefore correcting for multiple testing.

A few participants discussed which group was disadvantaged in their free-text justifications. In total, ten of the 502 participants justified their choice of one model over the other in terms of the disadvantaged group. For example, P120 (FP-Acc-Maj) wrote, "I would choose [the model with higher accuracy for Whites] since I am a white American and that model has a higher success rate for white Americans." As P439 (FP-Outcome-Maj) wrote, the model "disproportionately impacts white people. On the other hand, the whole bail system disproportionately impacts Black folks, so it may be a wash." Overall, 4.7% of participants volunteered in one of their free-text justifications that they considered the justice system as a whole to be unjust.

5 DISCUSSION

After discussing our protocol's limitations (Sec. 5.1), we compare our findings with those of prior empirical studies on perceptions of fairness (Sec. 5.2). We conclude by recapitulating our work's key lessons and proposing future work (Sec. 5.3).

5.1 Limitations

Our experiment was limited in a number of ways. First, we studied a convenience sample that is not necessarily generalizable to any larger group. We asked about differences in how models treated White and African-American defendants, but did not have a sufficiently large number of non-White participants to meaningfully determine whether there was an interaction between the participant's own demographics and what group was disadvantaged. In addition, we only examined fully automated models and fully manual human judges, whereas a hybrid approach (a human who relies in part on an automated model) is a potential compromise. Furthermore, we investigated only a single visualization of the model properties and differences. While we piloted these visualizations using cognitive interviews to verify their intelligibility for participants without particular statistics expertise, other work has used different visualizations, including simple text statements of percentages, tables, and novel visualizations. We used straightforward bar graphs, annotated with the text percentage. We chose bar graphs because they communicate difference and magnitude.

Because we used a convenience sample, the precise percentages reported are unlikely to generalize. However, we expect the broad patterns we saw to generalize, with implications for the design of fairness interventions. We found that even outside of the ethics and machine learning community, there exist strong, nuanced views about the acceptability of different approaches to fairness. Additionally, participants expressed significant differences of opinion.

5.2 Comparison with other empirical studies

As noted in the previous section, prior empirical studies of how humans perceive the fairness of machine learning models vary from each other, and from our work, in how they visualize the properties of these models. Our supplementary appendix discusses this confound further. While this limits our ability to directly compare with prior work, in this section we highlight similarities and differences

in conclusions among these studies. Future work could investigate how the visualization of models impacts perceptions of fairness.

Grgić-Hlača et al. found that the fairness of explicitly considering race to predict recidivism risk depended on how doing so impacted model accuracy [10]. While only 21% of their participants thought it unconditionally fair to consider race, 42% thought it fair to consider race if it improved accuracy. We investigated the fairness of explicitly considering race in a model that equalized accuracy across racial groups, rather than strictly increasing accuracy. While we observed high variance in fairness ratings for this model, most participants who rated the model as not at all fair mentioned the unfairness of using race, regardless of equalized accuracy. Among those who rated the model as fair, most referenced equalized accuracy as a mitigating factor (e.g., P12 wrote, "It's equally accurate for people of different races, so I think that makes the use of the data justified."). Our findings are therefore consistent with Grgić-Hlača et al.'s findings, suggesting that people sometimes find explicitly considering race justified if doing so improves performance.

Saxena et al. found the race of the person advantaged in loan allocation affected perceptions of fairness [24]. Specifically, participants found it more fair to give the entire sum to the candidate with the higher loan repayment rate than to divide it equally, but only when the candidate with the higher repayment rate was Black. In contrast, we did not observe any statistically significant effects when we compared conditions in which Whites were negatively impacted to those in which African-Americans were negatively impacted. This suggests that race may be more significant when examining individuals (as in Saxena et al.), rather than groups.

Srivastava et al. found a preference for demographic parity (equalizing the percentage of people who receive a positive classification, which was our = *Outcome* condition) over other definitions of fairness, like equalized false positive rates or false negative rates [26]. In contrast, we found that equalizing the false positive rate at the expense of having disparate outcome rates across groups was preferred over the opposite. One possible explanation for this lies in the different ways the two studies visualized the properties of a model, which we discuss further in the supplementary appendix.

Kennedy et al. found the size of the training data, the false positive and false negative rates, and the institutional source most impacted which model participants trusted [15]. We also found that equalizing false positive rates was generally valued over equalizing accuracy. Whereas Kennedy et al. found that their participants generally expressed trust in algorithmic methods, our participants expressed a general preference for a human judge. A possible explanation for this difference is that while we showed differences between model properties by racial group, Kennedy et al. investigated only the overall false positive rate, false negative rate, and accuracy. Had their participants been aware of differences across racial groups, they may have been less likely to trust algorithms.

5.3 Conclusions and Future Work

Our survey-based experiment asked participants to comparatively evaluate two models that exemplified the two sides of the realistic trade-offs between fairness-related properties. We observed a marginal, yet statistically significant, preference for equalizing false positives across demographic groups over equalizing accuracy. For other trade-offs, we observed at most a weak (and non-significant) preference. Notably, though, each side of each trade-off was strongly preferred by a non-trivial fraction of participants. This result casts doubt on the possibility of achieving broad acceptance across society that the *right* fairness decision was made among mutually exclusive, yet seemingly desirable, statistical definitions of fairness. Furthermore, we observed a general, yet not often not overwhelming, preference for a human judge over models capturing either side of the realistic trade-offs we examined. Even when participants thought that neither side of the trade-off was biased, over half of them still preferred a human judge over the model. We also found that just because a participant felt a model was "not at all biased" did not imply that they considered the model fair.

Our findings are a starting point for future investigations of algorithmic fairness from samples of participants without specialized knowledge or deep previous engagement with algorithmic justice. For example, a similar instrument could be used to understand how participants with first-hand experience of the criminal justice system approach these questions. Are the trade-offs similarly contentious? Do the explanations those participants provide differ? Though we did find a pattern of marginal support for equalization of false positives over accuracy, even in this a bare majority of participants supported one side over the other. Nonetheless, our results should not be read as an attempt to make definitive claims about which sides of trade-offs should be favored.

A machine learning developer confronted with the tough types of trade-off decisions we investigated might be tempted to crowdsource the decision by surveying the public. Our findings suggest that crowdsourcing is unlikely to produce consensus or full clarity about the decision. Deciding a trade-off on the basis of $50\% + \epsilon$ will likely leave a significant portion of people unhappy. Furthermore, artificially curtailing options to be just between models, rather than leaving options to not have a model at all, is also unlikely to elicit true preferences.

Future interventions should promote the visibility of design decisions. Data scientists are not well-situated to resolve the tensions and disagreements we have identified. Instead, data scientists should make clear the decisions they have made and allow experts and the public to deliberate about whether the model should be used. To that end, future work should investigate how to facilitate public involvement in decisions concerning fairness. In particular, future research into tools and mechanisms for identifying decisions during the data science workflow should be emphasized. Understanding the role the visualization of model properties plays in discourse about design decisions is key.

ACKNOWLEDGMENTS

This material is based upon work supported by the NSF Program on Fairness in AI in collaboration with Amazon under Award No. 1939728 (Title: "FAI: Identifying, Measuring, and Mitigating Fairness Issues in AI"). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or Amazon.

REFERENCES

[1] Philip Adler, Casey Falk, Sorelle A. Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing Black-box Models for Indirect Influence. *Knowledge and Information Systems* 54, 1 (Jan. 2018), 95–122.

[2] Julia Angwin and Jeff Larson. 2016. Machine Bias. ProPublica. (May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine Experiment. *Nature* 563, 7729 (Nov. 2018), 59–64.

[4] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "It's Reducing a Human Being to a Percentage"; Perceptions of Justice in Algorithmic Decisions. In *Proc. CHI.*

[5] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big data* 5, 2 (2017), 153–163. arXiv: 1610.07524.

[6] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *Proc. IEEE S&P.*

[7] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Northpointe Incorporated. (2016). http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

[8] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2014. Certifying and Removing Disparate Impact. arXiv. (2014). http://arxiv.org/abs/1412.3756.

[9] Nina Grgić-Hlača, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proc. WWW.*

[10] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *Proc. NIPS Symposium on Machine Learning and the Law.*

[11] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *Proc. AAAI.*

[12] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proc. NIPS.*

[13] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proc. CHI.*

[14] Farnaz Jahanbakhsh, Wai-Tat Fu, Karrie Karahalios, Darko Marinov, and Brian Bailey. 2017. You Want Me to Work with Who?: Stakeholder Perceptions of Automated Team Formation in Project-based Courses. In *Proc. CHI.*

[15] Ryan Kennedy, Philip Waggoner, and Matthew Ward. 2018. Trust in Public Policy Algorithms. SSRN. (2018). https://papers.ssrn.com/abstract=3339475.

[16] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Sch olkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Proc. NIPS.*

[17] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv. (2016). http://arxiv.org/abs/1609.05807.

[18] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. (May 2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[19] Arvind Narayanan. 2018. Tutorial: 21 fairness definitions and their politics. (2018). https://www.youtube.com/watch?v=jIXIuYdnyyk.

[20] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. 2019. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *Proc. IEEE S&P.*

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. In *Proc. KDD.*

[22] Aaron Roth. 2018. Aaron Roth on Twitter: "A nice piece by @juliapowles and @HNissenbaum reminding us that societal injustice is complicated, whereas technical papers on 'fairness in machine learning' are necessarily simplistic and narrow. https://t.co/J6k9Mz6ELY But let me take a moment to defend technical FATML work. 1/" / Twitter. (2018). https://twitter.com/aaroth/status/1071480797306257408

[23] Chris Russell. 2019. Efficient Search for Diverse Coherent Explanations. In *Proc. FAT*.*

[24] Nripsuta Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In *Proc. AIES.*

[25] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proc. FAT*.*

[26] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proc. KDD.*

[27] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proc. FAT*.*

[28] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proc. CHI.*

[29] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment. In *Proc. WWW.*

APPENDIX

A ADDITIONAL FIGURES

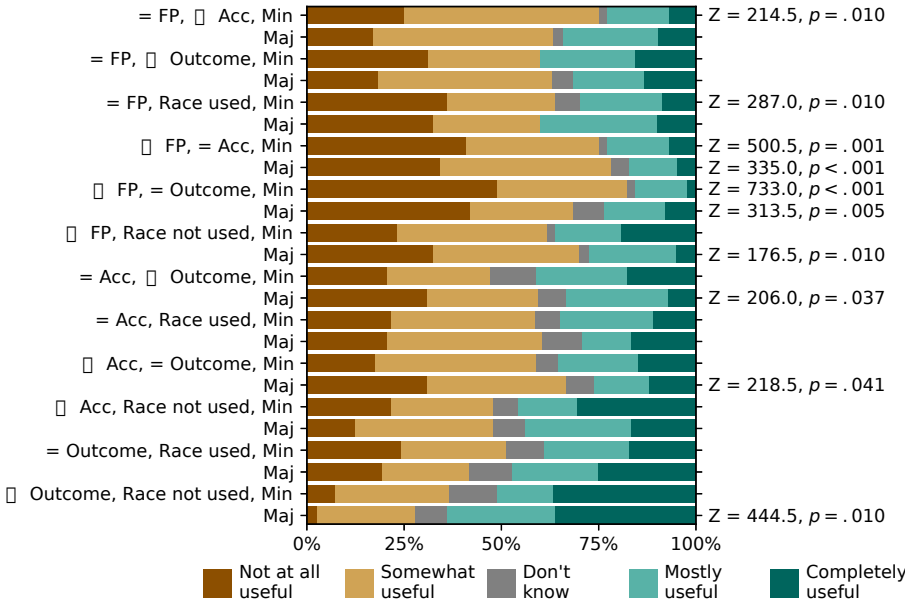

Figure 7: Participants' ratings of the usefulness of each model.

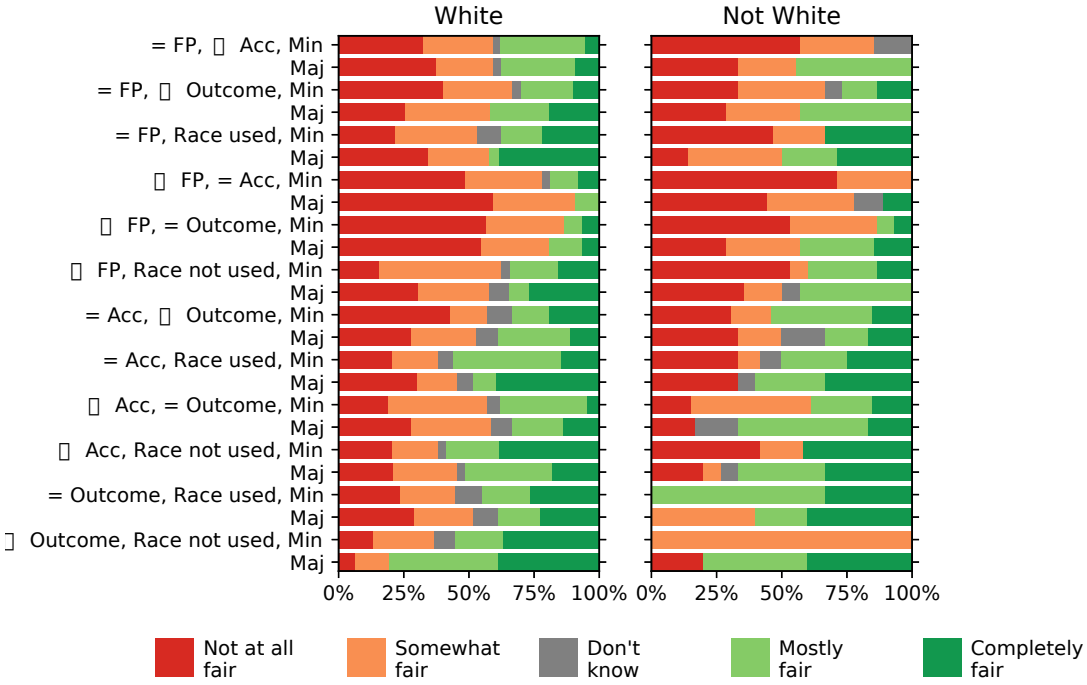

Figure 8: Fairness ratings of models broken out by whether participant identified as white.

B DEMOGRAPHIC MODELS

	Bias	Fairness	Usefulness
Income (linear fit)	−1.10 (1.31)	1.11 (1.17)	0.10 (1.18)
Education (linear fit)	1.73 (1.01)	−0.76 (0.91)	0.29 (0.93)
Tech experience (linear fit)	−0.19 (0.24)	0.33 (0.24)	0.05 (0.22)
Race: Black	0.37 (0.23)	−0.26 (0.23)	−0.11 (0.23)
Race: Asian	0.05 (0.24)	−0.25 (0.24)	0.04 (0.23)
Race: Latinx	−0.12 (0.29)	0.22 (0.30)	−0.40 (0.30)
Race: Prefer not to say	1.37 (0.71)	−0.77 (0.71)	−0.44 (0.66)
Political: Independent	−0.21 (0.16)	0.11 (0.15)	0.01 (0.15)
Political: Other	−0.30 (0.50)	−1.03 (0.51)*	−1.60 (0.57)**
Political: Prefer not to say	−0.96 (0.77)	1.35 (0.77)	0.07 (0.73)
Political: Republican	−0.17 (0.17)	0.39 (0.16)*	0.11 (0.17)
Gender: Female	−0.01 (0.13)	−0.16 (0.13)	0.05 (0.13)
Log Likelihood	-1337.95	-1325.11	-1332.19
AIC	2761.90	2736.23	2750.39
BIC	2969.35	2943.68	2957.83
Num. obs.	920	920	920
Groups (trade-off)	24	24	24
Variance: trade-off: (Intercept)	0.22	0.29	0.21

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 3: Random effects model coefficients for model judgments. The standard error for each coefficient is reported in parentheses. We treated income, education, and tech experience as ordinal variables (fit linearly) with the lowest income, educational background, and amount of tech experience as the baselines. We treated race, political affiliation, and gender as categorical, with White, Democrat, and Male as the baselines.

C SURVEY INSTRUMENT

Introduction

Thank you for choosing to participate in this study. After a short tutorial about machine learning, you will be asked to react to a scenario. After this, we will ask you some demographic questions to better understand your responses. In this scenario you will be asked to play the role of a data scientist. Specifically you will be asked to use your judgment about which model to use for a particular task.

- Machine learning models can help you use past data to make predictions.
  - *Training Data*: To make a model that predicts if a college basketball team will make the tournament, you could first gather data from past seasons about the characteristics, or features of teams that did or did not make the tournament.
- Machine learning models use *patterns* in the *training data* to make predictions.
  - For example, it might find that the coach's number of years of experience combined with number of seniors on the team best predict whether the team will make the playoffs.
  - The collection of patterns is the *model*.
- A model might not make the right prediction for every team.
  - For example, a team with an experienced head coach and many seniors might also have an injured star player or might just have bad luck, and not make the playoffs.
- Once the model has been made, it is possible to test how it will perform by applying it to data that was set aside before building the model.
  - For example, we could make the model using data from the 2015 and 2016 seasons, and see how well it works in the 2017 season.

Start of New Scenarios

The next series of questions will refer to this scenario. Metropolis city needs to decide to which people charged with non-violent offenses they can grant bail (and thus potentially release from jail) pending trial.

- *Training Data*: The city created two models by gathering data from past non-violent offense charges about the characteristics of people that did or did not show up for their trial.

- *Models*: The resulting models try to predict whether a person newly charged with a non-violent offense will or will not show up for their trial.

Based on the model Metropolis city chooses, they can decide which people to release pending trial:

- If the model predicts that a person **will** show up for their trial, they will be granted bail and potentially released.
- If the model predicts that a person **will not** show up for their trial, they will be denied bail, and remain in jail.

However, neither model they are considering is perfect. They each make mistakes in different ways.

Model X [Order of model X and model Y randomized]. Model X is one of the models Metropolis City is considering using. Below are two graphs showing properties of model X. The top graph shows the [**accuracy**] of model X. The bottom graph shows the [**probability of being granted bail**] when model X is used.

- if **Acc** a quality: **Accuracy**: the accuracy of the model is the rate at which the model makes correct predictions. A prediction is correct if the model either predicts that a defendant will show up for their trial and they would, or that they will not show up for their trial and they would not have
- if **Outcome** a quality: **Probability of bail**: the probability of bail is the likelihood of a defendant being granted bail if the model is used
- if **FP** a quality: **Mistakenly denied bail**: A defendant who is *mistakenly denied bail* is one that the model predicts would not show up for their trial when they would have
- if **Race usage** a quality: **Race is not one of the features used**: the model makes decisions with no knowledge of the race of the subjects. Other features like type of offense and number of previous offenses are used as input to the model.

(1) Do you think model X is fair? (Not at all fair, Somewhat fair, Mostly fair, Completely fair, Don't know)
(2) Why?
(3) Do you think model X is biased (Not at all biased, Somewhat biased, Mostly biased, Completely biased, Don't know)
(4) Why?
(5) Do you think model X is useful (Not at all useful, Somewhat useful, Mostly useful, Very useful, Don't know)
(6) Why?
(7) Given a choice between model X and a human judge to make bail decisions, what would you choose? (Definitely model X, Probably model X, Unsure/can't decide, Probably human judge, Definitely human judge)
(8) Why?

Model Y. Model Y is one of the models Metropolis City is considering using. Below are two graphs showing the properties of model Y. The top graph shows the [**percent probability of being granted bail**] when model Y is used. The bottom graph shows what percent of defendants are [**mistakenly denied bail**] when model Y is used.

- *Descriptions the same as in question about model X*
- *Questions the same as for model X*

Changed Answers?

(1) Did seeing the second model change the answers you wished to give for the first model you saw? (Yes, No)
(2) What aspects of your model would you wish to change and why?

Model Comparisons [Figure showing model X next to model Y].

(1) Which model is more fair, model X or model Y? (Definitely model X, Probably model X, Models X and Y are equally fair, Probably model Y, Definitely model Y)
(2) Why?
(3) Which model is more biased, model X or model Y? (Definitely model X, Probably model X, Models X and Y are equally biased, Probably model Y, Definitely model Y)
(4) Why?
(5) Which model is more useful, model X or model Y? (Definitely model X, Probably model X, Models X and Y are equally useful, Probably model Y, Definitely model Y)
(6) Why?
(7) Given a choice between model X and model Y, which would you choose? (Definitely model X, Probably model X, Unsure/can't decide, Probably model Y, Definitely model Y)
(8) Why?
(9) Given a choice between using a model (either model X or model Y) or a human judge to make the decision, what would you choose? (Definitely a model, Probably a model, Unsure/can't decide, Probably human judge, Definitely human judge)
(10) Why?

Graph Understanding. Below is a question relating to the following graph. **Do not use information you may have seen in other graphs in answering this question**.

[if accuracy a trade-off, figure showing a Group A having 79% accuracy and Group B having 53% accuracy]
Please select which statement(s) can be validly inferred from this graph.

(1) More people in group A will receive bail than people in group B
(2) Predictions about people in group A will be more likely to be correct than predictions about people in group B
(3) People in group A will be more likely to be mistakenly denied bail than people in group B
(4) People in group A will be more likely to be mistakenly granted bail than people in group B

[if FP a trade-off, figure showing a Group A having 24% false positives and Group B having 45% accuracy]
Please select which statement(s) can be validly inferred from this graph.

(1) More people in group A will receive bail and not show up than people in group B
(2) Predictions about people in group B will be more likely to be mistakenly denied bail than people in group A
(3) People in group A will be more likely to be mistakenly granted bail than people in group B
(4) People in group A will be more likely to receive bail than people in group B

[if outcome a trade-off, figure showing a Group A having 65% bail probability and Group B having 41% bail probability]
Please select which statement(s) can be validly inferred from this graph.

(1) More people in group A will receive bail than people in group B
(2) People in group A will be more likely to receive bail than people in group B
(3) Predictions about people in group A will be more likely to be correct than predictions about people in group B
(4) People in group B will be more likely to be mistakenly denied bail than people in group A

Cognitive Reflection Test.

(1) A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?
(2) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?
(3) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half the lake?

Demographic Questions. You will now be asked a series of demographic questions.

(1) What is your age (18-24 years old, 25-34 years old, 35-44 years old, 45-54 years old, 55-64 years old, 65+ years old, prefer not to say)
(2) What is your gender (Male, Female, Gender non-binary, Other not listed, prefer not to say)
(3) Please select the set of categories that describe your racial or ethnic background. You may select multiple categories (American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Pacific Islander, White, prefer not to say)
(4) What is your highest level of formal education? (Less than high school, High school graduate, some college, 2 year degree, 4 year degree, Professional degree, Doctorate, prefer not to say)
(5) Generally speaking do you think of yourself as a Republican, Democrat, Independent or something else? (Republican, Democrat, Independent, Other, prefer not to say)
(6) What is your annual household income? (Less than $20,000,$20,000 to $49,999, $50,000 to $99,9999, $100,000 to $249,999, Over $250,000, prefer not to say)
(7) What is your experience with computer science? Please select as many or as few options that apply.
    - I have taken a computer science course
    - I have taken a course where computer science was mentioned as a topic or read about computer science topics online
    - I have or had a job where computer science tasks were part of my job duties (I have written, documented, or manipulated code)
    - I have never received computer science education or held a computer science job
(8) What is your experience with machine learning? Please select as many or as few options that apply.
    - I have taken a course in machine learning
    - I have taken a course where machine learning was mentioned as a topic or read about machine learning topics online
    - I have or had a job where machine learning tasks were part of my job duties (model training, model debugging etc)
    - I have never received machine learning education or held a job in which it was used
(9) What is your experience with probability? Please select as many or as few options that apply.
    - I have taken a course in probability
    - I have taken a course where probability was mentioned as a topic or read about probability topics online
    - I have or had a job where probability related tasks were part of my job duties
    - I have never received probability education or held a job in which it was used

D COMPARISON OF MODEL VISUALIZATIONS WITH OTHER EMPIRICAL WORK

In work by Srivastava et al. [26] participants were asked to make twenty comparisons generated through an adaptive algorithm meant to converge upon each participant's preferred notion of fairness. First, we note that the examples generated under this methodology do not necessarily encode any notion of trade-off between properties. Second, the way in which the information was displayed may play a role in their finding of a preference for demographic parity. The study used a display showing stylized pictures of people with different combinations of races and genders along with the true label. Below this were two rows each containing a color coded prediction for each person (see Figure 9). We hypothesize that this display may require a greater amount of cognitive load to compare quantities like accuracy or false positive rates, relative to demographic parity. In order to calculate such a quantity, the survey taker would need to count and remember the number of misclassifications by group. Then the survey taker would need to calculate the rate between groups. Demographic parity is much more visually clear. A survey taker could count the number of positive classifications for each group. Since there were ten people displayed, an approximate count would suffice to move towards demographic parity. Srivastava et al. also performed a survey where they asked survey respondents to choose between three models with differing overall and intergroup accuracy. However in this experiment, they did not test qualities against one another.

By contrast, our way of depicting differences between models encodes far less information (see Figure 10). At the same time, it is much more clear about the differences we are trying to test.

At the time of publication, visual depictions of models used by Kennedy et al. [15] were unavailable.

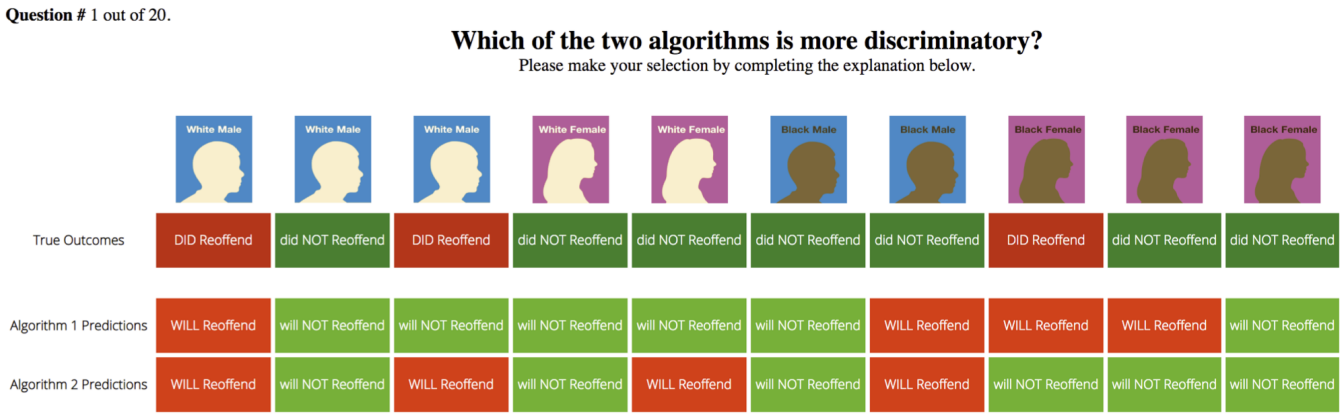

Figure 9: Visualization used by Srivastava et al. [26] in their investigation of fairness considerations.

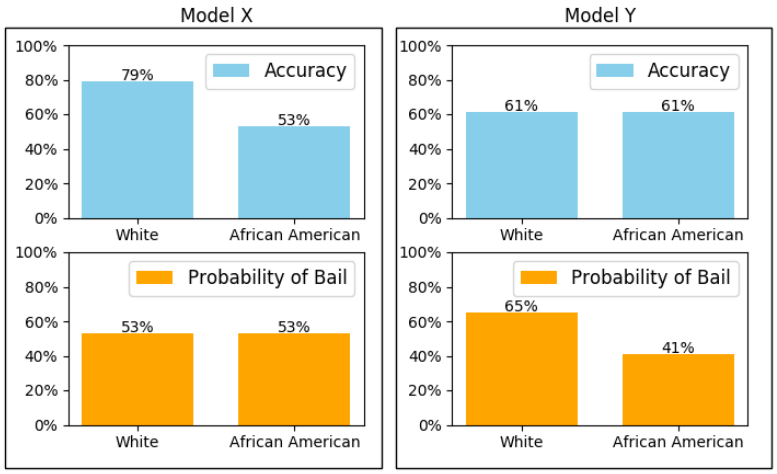

Figure 10: A visualization of two models from our study. This one depicts the two models in the Acc-Outcome-Min condition.